

# **Эконометрика**

## **Лекция 1, часть 2**

### **Повторение теории вероятностей и математической статистики**

**Демидова О.А.**

**[https://www.hse.ru/staff/demidova\\_olga](https://www.hse.ru/staff/demidova_olga)**

**E-mail:demidova@hse.ru**

**2020**

# Теория вероятностей. Случайные величины

---

Опр. Случайными величинами называют числовые функции,  
определенные на множестве элементарных событий:

$$X : \Omega \rightarrow R$$

# Теория вероятностей. Дискретные случайные величины

---

**Опр.** Если случайная величина принимает конечное или счетное множество значений, то она называется дискретной.

**Дискретные случайные величины удобно задавать с помощью таблицы,**  
**в первой строке которой перечислены значения, которые принимает случайная величина, а во второй – соответствующие вероятности:**

<b>X</b>	<b><math>X_1</math></b>	<b>...</b>	<b><math>X_n</math></b>
<b>P</b>	<b><math>P_1</math></b>	<b>...</b>	<b><math>P_n</math></b>

# Пример дискретной случайной величины

---

Случайная величина  $X$  — количество очков на верхней грани брошенной кости

$X$	1	2	3	4	5	6
$P$	1/6	1/6	1/6	1/6	1/6	1/6

# Функция распределения случайной величины

---

Опр. Функцией распределения  $F_X(x)$  случайной величины  $X$  называется  $F_X(x) = P(X \leq x)$ .

Свойства функции распределения:

1) 
$$\lim_{x \rightarrow -\infty} F(x) = 0$$

2) 
$$\lim_{x \rightarrow \infty} F(x) = 1$$

3)  $F(x)$  – неубывающая функция

4)  $F(x)$  является непрерывной справа

# Непрерывная случайная величина

---

**Опр.** Случайная величина называется непрерывной, если существует кусочно непрерывная функция  $f(x)$  такая, что  $F'(x) = f(x)$ .

$f(x)$  называется функцией плотности распределения.

**Свойства функции плотности**

$$1) \quad f(x) \geq 0$$

$$2) \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

$$3) \quad P(a \leq X \leq b) = \int_a^b f(x) dx$$

# Математическое ожидание случайной величины

---

Существует две основных числовых характеристики случайных величин: математическое ожидание и дисперсия.

Опр. Математическое ожидание случайной величины:

$$E(X) = \sum_{i=1}^n X_i p_i \quad , \text{если } X - \text{дискретная случайная величина,}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad , \text{если } X - \text{непрерывная случайная величина.}$$

# Дисперсия случайной величины

---

**Опр. Дисперсией (обычно обозначаемой  $\sigma^2$ ) случайной величины называется:**

$$\text{Var}(X) = \sigma_X^2 = E(X - E(X))^2 .$$

**Опр. Стандартным отклонением называется корень из дисперсии.**



# Ковариация и коэффициент корреляции случайных величин X и Y

---

Опр. Ковариацией случайных величин X и Y называется

$$\text{Cov}(X, Y) = E(X - E(X))(Y - E(Y))$$

Опр. Коэффициентом корреляции случайных величин X и Y называется:

$$r_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Свойства коэффициента корреляции:

- 1)  $|r_{XY}| \leq 1$
- 2) Если  $r_{XY} = 0$ , то не существует линейной связи между X и Y
- 3) Если  $|r_{XY}| = 1$ , то между случайными величинами X и Y существует точная линейная связь:  $Y = aX + b$

## Свойства математического ожидания, дисперсии и ковариации

---

1)  $E(C) = C$

2)  $E(CX) = CE(X)$

3)  $E(X + Y) = E(X) + E(Y)$

4)  $Var(C) = 0$

5)  $Var(CX) = C^2Var(X)$

6)  $Var(X + Y) = Var(X) + 2 Cov(X, Y) + Var(Y)$

7)  $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$

8)  $Cov(CX, Y) = CCov(X, Y)$

9)  $Cov(X, Y) = Cov(Y, X)$

10)  $Cov(X, X) = Var(X),$

11)  $Cov(X, C) = 0$

где  $C$  - константа,  $X, Y, Z$  – случайные величины .

# Совместное распределение двух случайных величин

---

Пусть  $X, Y$  - случайные величины с совместным законом распределения.

Это может быть таблица, если  $X, Y$  принимают конечное или счетное множество значений. Закон совместного распределения непрерывных случайных величин может быть задан с помощью совместной функции плотности  $f(x, y)$ .

# Маргинальные распределения

---

Если задан совместный закон распределения случайных величин  $X$  и  $Y$ , то маргинальное распределение случайной величины  $X$  имеет вид:

$P(X = X_i) = \sum_j P(X = X_i, Y = Y_j)$ ,  $i = 1, \dots, n$  для дискретного случая,

$f_x(x) = \int f(x, y) dy$  – функция плотности для непрерывной случайной величины.

Математическое ожидание и дисперсия случайных величин  $X$ ,  $Y$  определяются как обычно.

# Условные распределения

---

Условная плотность распределения определяется следующим образом:

$P(Y = Y_j | X = X_i) = P(X = X_i, Y = Y_j) / P(X = X_i)$  в дискретном случае,

$f(y|x) = f(x,y)/f_x(x)$  в непрерывном случае.

# Независимость случайных величин

---

**Если**

**$P(Y = Y_j | X = X_i) = P(Y = Y_j)$  для всех  $i$  в дискретном случае,  
или  $f(y|x) = f(y)$  в непрерывном случае, то случайные  
величины  $X, Y$  называются независимыми.**

**В случае независимости случайных величин  $X, Y$**

**$P(X = X_i, Y = Y_j) = P(X = X_i) P(Y = Y_j)$  в дискретном случае,  
 $f(x,y) = f_x(x) f_Y(y)$  в непрерывном случае.**

## Условное математическое ожидание

---

### Условное математическое ожидание

$E(Y|X = X_i) = \sum_j Y_j P(Y = Y_j|X = X_i)$  в дискретном случае,

$E(Y|X) = \int yf(y|x)dy$  в непрерывном случае.

# Нормальное распределение

---

**Опр. Случайная величина  $X$  имеет нормальное распределение с математическим ожиданием  $a$  и дисперсией  $\sigma^2$ , сокращенно это обозначается**

$$X \sim N(a, \sigma^2)$$

**если функция плотности этой случайной величины имеет**

**вид**

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x-a)^2}{2\sigma^2} \right\}$$



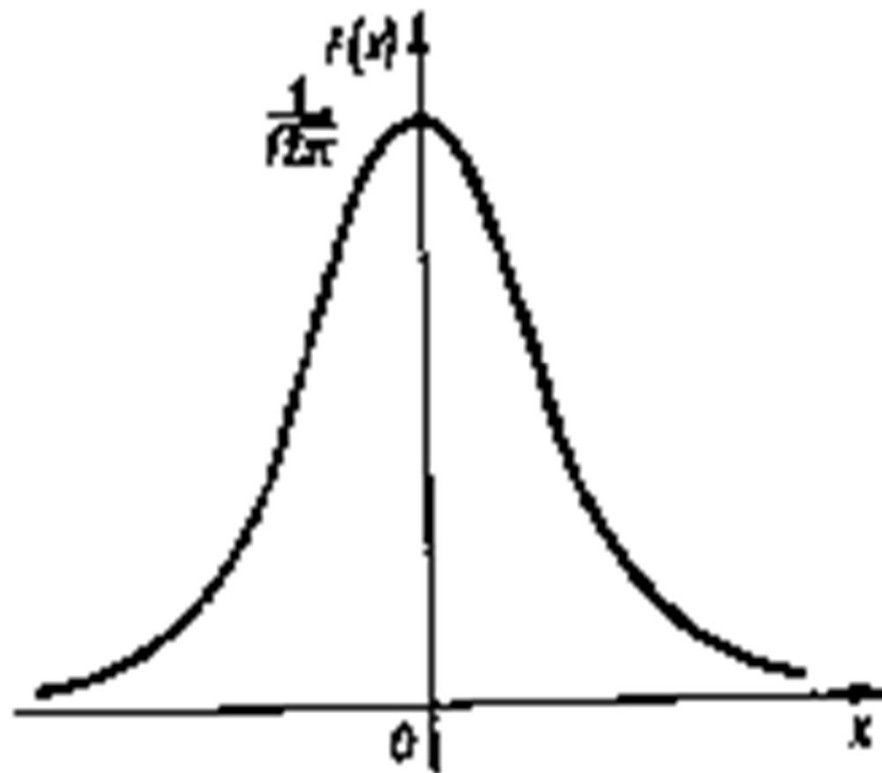
# Нормальное распределение

---

**Опр. Случайная величина  $X$  имеет стандартное нормальное распределение, если**

$$X \sim N(0,1)$$

# Функция плотности нормально распределенной случайной величины



## “Хи - квадрат” распределение

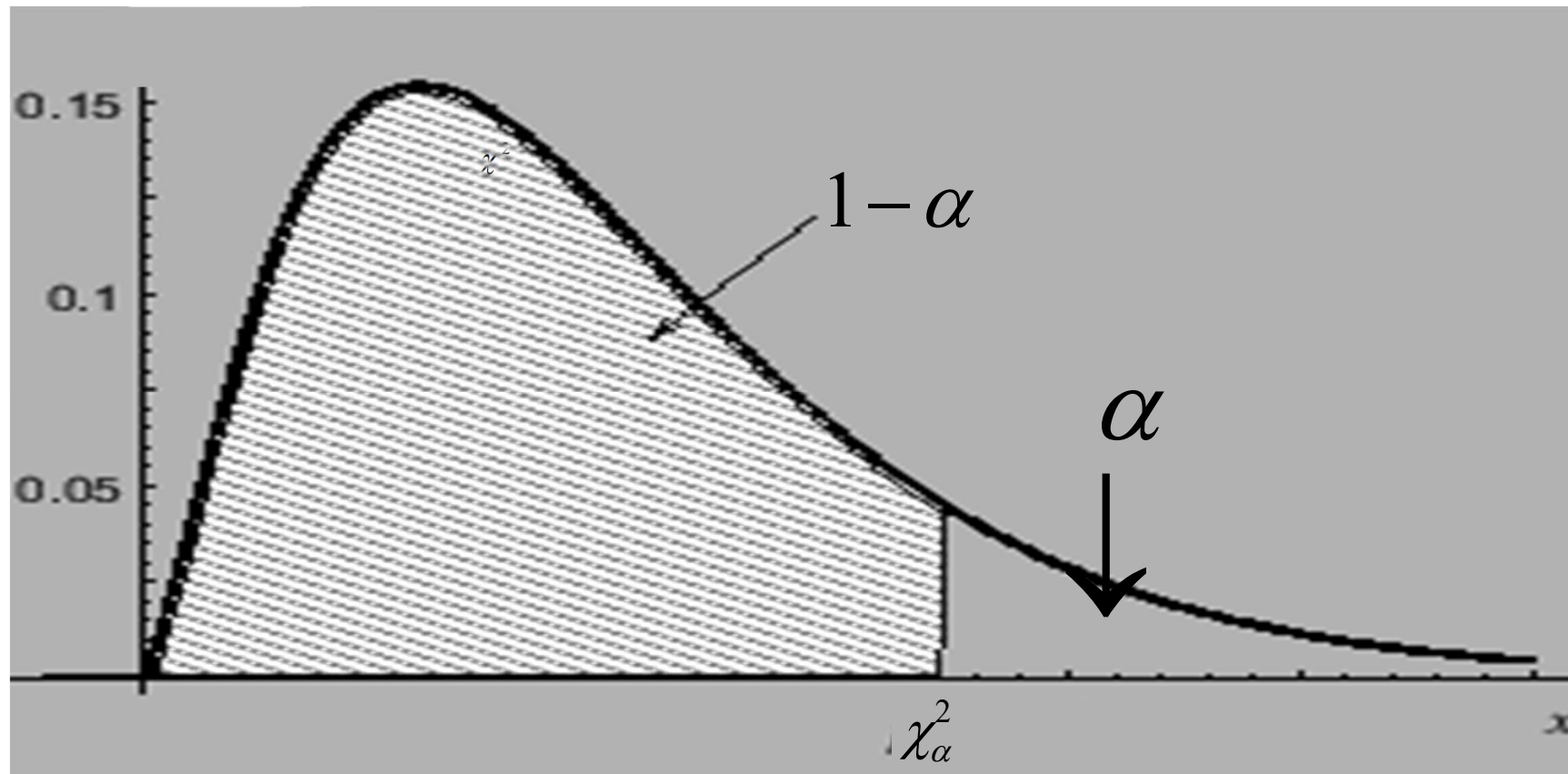
---

Опр. Случайная величина  $Y$  имеет “Хи – квадрат” распределение с  $k$  степенями свободы (сокращенно  $Y \sim \chi^2(k)$ ),

если  $Y = X_1^2 + \dots + X_k^2$ ,

где случайные величины  $X_i$  – независимые нормально распределенные случайные величины с математическим ожиданием 0 и дисперсией 1.

# Функция плотности распределения “Хи – квадрат”



# Таблицы для “Хи – квадрат” распределения

$\chi^2$  (хи-квадрат) распределение:

Критические значения  $\chi^2$

Уровень значимости	5%	1%	0.1%
Число степеней свободы			
1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.345	16.266
4	9.488	13.277	18.467
5	1.070	15.086	20.515
6	12.592	16.812	22.458
7	4.067	18.475	24.322
8	15.507	20.090	26.124
9	16.919	21.666	27.877
10	18.307	23.209	29.588

## t - распределение

---

**Опр. Случайная величина  $Z$  имеет  $t$  – распределение с  $k$  степенями свободы (сокращенно  $Z \sim t(k)$ ),**

**если**

$$Z = \frac{X}{\sqrt{Y / k}},$$

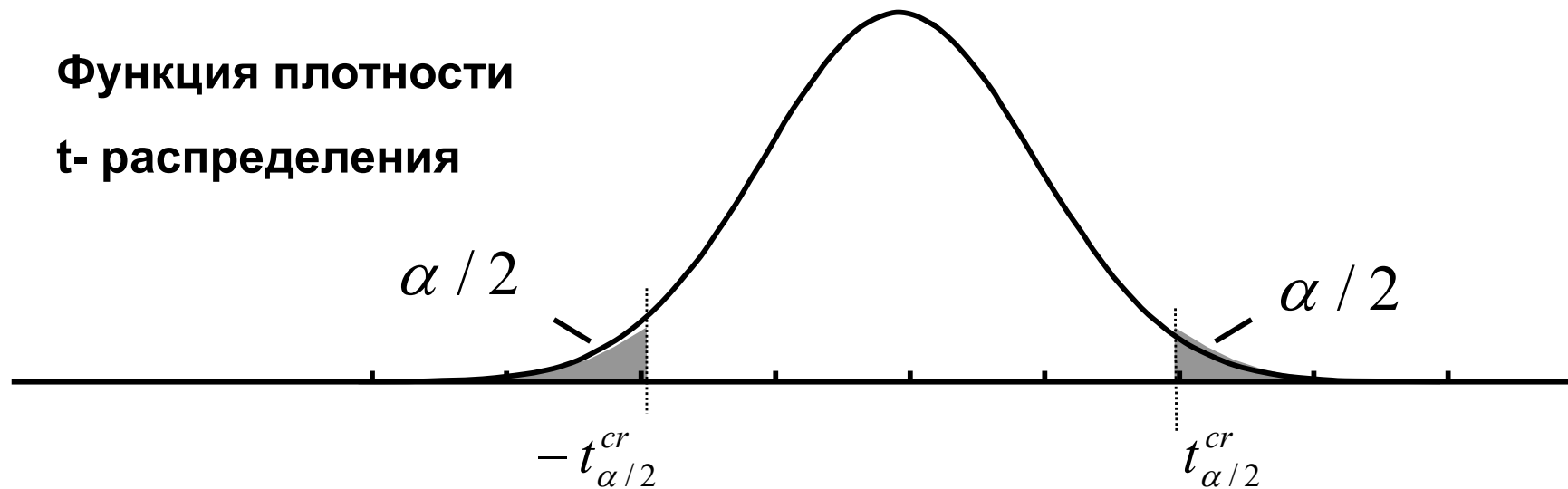
**где  $X \sim N(0,1)$ ,  $Y$  имеет “хи – квадрат” распределение с  $k$  степенями свободы,  $X$  и  $Y$  независимы.**

## **$t$ - распределение: Критические значения $t$**

Число степеней свободы	Двусторонний тест		10%	5%	2%	1%	0.2%	0.1%
	Односторонний тест		5%	2.5%	1%	0.5%	0.1%	0.05%
1			6.314	12.706	31.821	63.657	318.31	636.62
2			2.920	4.303	6.965	9.925	22.327	31.598
3			2.353	3.182	4.541	5.841	10.214	12.924
4			2.132	2.776	3.747	4.604	7.173	8.610
5			2.015	2.571	3.365	4.032	5.893	6.869
...		...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
18			1.734	2.101	2.552	2.878	3.610	3.922
19			1.729	2.093	2.539	2.861	3.579	3.883
20			1.725	2.086	2.528	2.845	3.552	3.850
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
120			1.658	1.980	2.358	2.617	3.160	3.373
			1.645	1.960	2.326	2.576	3.090	3.291

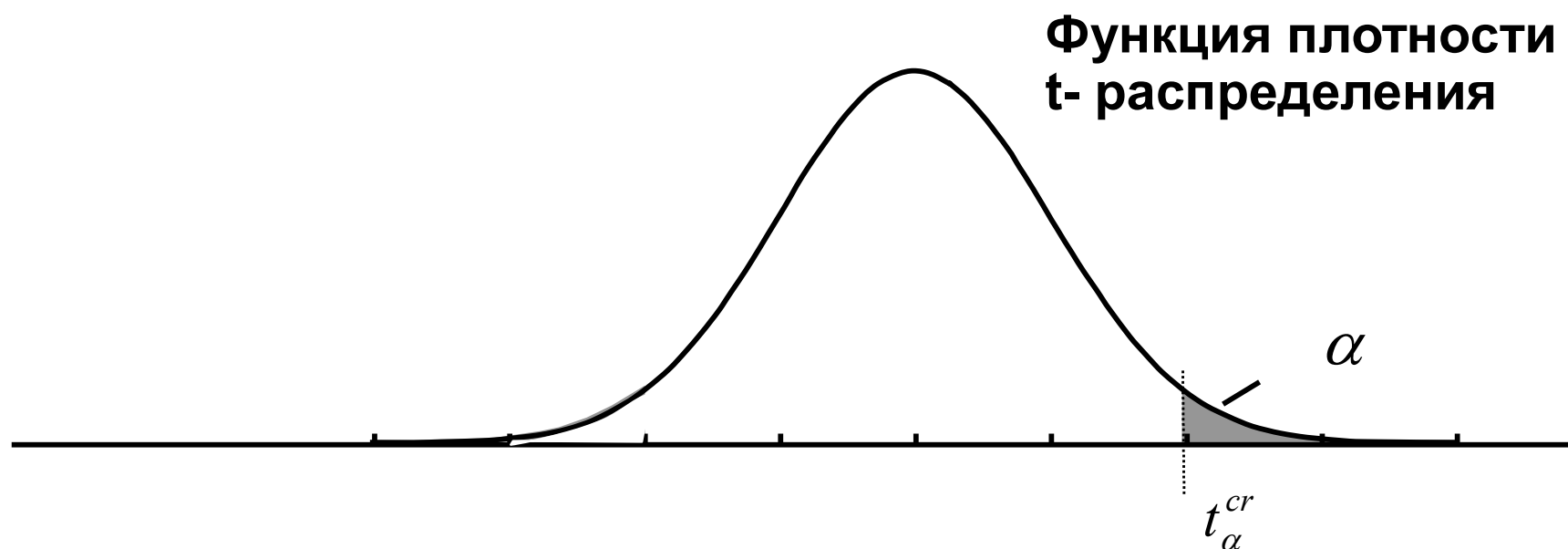
# Функция плотности $t$ - распределения. Двусторонний тест

Функция плотности  
 $t$ - распределения





## Функция плотности $t$ - распределения. Односторонний тест



## F - распределение

---

Опр. Случайная величина  $Z$  имеет F - распределение со степенями свободы  $m$  и  $n$  (сокращенно  $Z \sim F(m, n)$ ),

$$\text{если } Z = \frac{X / m}{Y / n},$$

где случайная величина  $X$  имеет распределение “хи– квадрат” с  $m$  степенями свободы, случайная величина  $Y$  имеет распределение “хи– квадрат” с  $n$  степенями свободы,  $X$  и  $Y$  независимы.

# F - распределение

F -распределение: Критические значения F (5% уровень значимости)

$v^1$	25	30	35	40	50	60	75	100	150	200
$v^2$										
1	249.26	250.10	250.69	251.14	251.77	252.20	252.62	253.04	253.46	253.68
2	19.46	19.46	19.47	19.47	19.48	19.48	19.48	19.49	19.49	19.49
3	8.63	8.62	8.60	8.59	8.58	8.57	8.56	8.55	8.54	8.54
4	5.77	5.75	5.73	5.72	5.70	5.69	5.68	5.66	5.65	5.65
5	4.52	4.50	4.48	4.46	4.44	4.43	4.42	4.41	4.39	4.39
6	3.83	3.81	3.79	3.77	3.75	3.74	3.73	3.71	3.70	3.69
7	3.40	3.38	3.36	3.34	3.32	3.30	3.29	3.27	3.26	3.25
8	3.11	3.08	3.06	3.04	3.02	3.01	2.99	2.97	2.96	2.95
9	2.89	2.86	2.84	2.83	2.80	2.79	2.77	2.76	2.74	2.73
10	2.73	2.70	2.68	2.66	2.64	2.62	2.60	2.59	2.57	2.56
11	2.60	2.57	2.55	2.53	2.51	2.49	2.47	2.46	2.44	2.43
12	2.50	2.47	2.44	2.43	2.40	2.38	2.37	2.35	2.33	2.32

# Математическая статистика

---

**Совокупность всех возможных значений случайной величины называется генеральной совокупностью. Подмножество генеральной совокупности называется выборкой.**

**Основная задача математической статистики – оценивание характеристик генеральной совокупности по выборке.**

**Обо всей генеральной совокупности мы, как правило, ничего не знаем точно и можем строить лишь догадки - гипотезы. Для проверки своих гипотез мы исследуем независимую выборку из генеральной совокупности и строим на основании выборки выборочные оценки неизвестных теоретических параметров.**

**Различают точечные и интервальные оценки.**

## Точечные оценки

---

Предположим, что мы имеем выборку  $X_1, \dots, X_n$  из распределения, зависящего от параметра  $\theta$ .

Опр. Точечной оценкой (статистикой) называется любая числовая функция от выборки  $\hat{\theta}(X_1, \dots, X_n)$ .

# **Несмещенность, эффективность, состоятельность оценок**

---

**Точечные оценки считаются «хорошими», если они обладают определенными свойствами:**

- несмещенностью (в этом случае математическое ожидание оценки совпадает с оцениваемым теоретическим параметром);**
- состоятельностью (это означает, что для больших выборок вероятность значимых отклонений величины оценки от значения оцениваемого теоретического параметра равна нулю);**
- эффективностью (чем меньше дисперсия оценки, тем она считается эффективнее).**

## Несмещенные оценки для математического ожидания и дисперсии

---

Предположим,  $X_1, \dots, X_n$  - выборка из генеральной совокупности,  $E(X_i) = \mu$ ,  $D(X_i) = \sigma^2$ ,  $i = 1, \dots, n$ .

Несмещенные оценки для математического ожидания и дисперсии (выборочное среднее и выборочная дисперсия) :

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

## Несмещенная оценка для ковариации

---

Для двух выборок  $X_1, \dots, X_n$  и  $Y_1, \dots, Y_n$  несмещенная оценка для ковариации случайных величин  $X$  и  $Y$  имеет вид:

$$\hat{\text{cov}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$



# Интервальные оценки

---

При интервальном оценивании конструируются две функции от выборки:

$$\hat{\theta}_1(X_1, \dots, X_n) \quad \text{и} \quad \hat{\theta}_2(X_1, \dots, X_n)$$

такие, что

$$1 - \alpha = P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2)$$

Этот интервал называется  $(1 - \alpha)100\%$  доверительным интервалом для параметра  $\theta$ .

# Проверка гипотез

---

Предположим, что мы имеем выборку  $X_1, \dots, X_n$  из распределения, зависящего от параметра  $\theta$ .

Относительно параметра  $\theta$  выдвигаются две гипотезы, основная  $H_0$  и альтернативная  $H_1$ , например:

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

# Проверка гипотез

---

**Статистическим тестом (или просто тестом) называется процедура, основанная на наблюдениях  $X_1, \dots, X_n$ , результатом которой является одно из двух возможных решений:**

- 1) Не отвергать основную гипотезу  $H_0$ ,**
- 2) Отвергнуть нулевую гипотезу  $H_0$  в пользу альтернативной гипотезы  $H_1$ .**

**При этом можно совершить две ошибки:**

- 1) Ошибка первого рода – отвергнуть нулевую гипотезу, когда она верна,**
- 2) Ошибка второго рода – не отвергнуть нулевую гипотезу, когда она не верна.**

# Проверка гипотез

---

Вероятность ошибки первого рода обозначается  $\alpha$  и называется уровнем значимости теста,

Вероятность ошибки второго рода обозначается  $\beta$ .

$1 - \beta$  называется мощностью теста.

## Проверка гипотез

На практике для построения тестов часто используют следующий подход. Находят такую статистику  $t_n(X_1, \dots, X_n)$ , что если гипотеза  $H_0$  верна, то распределение случайной величины  $t_n$  известно. Тогда для заданного уровня значимости  $\alpha$  можно найти такую область  $K_\alpha$ , что  $P(t_n \in K_\alpha) = 1 - \alpha$ .

Тогда тест проводится следующим образом:

- 1) На основании наблюдений  $X_1, \dots, X_n$  вычисляется значение статистики  $t_n$ .
- 2) Для заданного уровня значимости  $\alpha$  находится область  $K_\alpha$ .
- 3) Если  $t_n \in K_\alpha$ , то нулевая гипотеза не отвергается.
- 4) В противном случае нулевая гипотеза отвергается в пользу альтернативной.

# Проверка гипотез

---

Статистику  $t_n$  называют критической статистикой, а область  $K_\alpha$  – критической областью.

На практике критические статистики часто имеют распределение  $N(0,1)$ ,  $t$ , «хи – квадрат»,  $F$ .

В этих случаях для критической статистики легко рассчитать  $p$ -value ( $p$ -значение) – минимальный уровень значимости, при котором основная гипотеза отвергается.